Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 59-66 recibido 20-01-10 revisado 19-02-10 aceptado 04-03-10

Localización de Palabras basada en Grafos de Fonemas*

Word Spotting based on Phoneme Graphs

Jon Ander Gómez Adrián, Marcos Calvo Lance, Emilio Sanchis Arnal
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera, s/n — 46022 Valencia (España)
{jon,mcalvo,esanchis}@dsic.upv.es

Resumen: En este artículo se propone la utilización de grafos de fonemas para tareas de detección y localización de palabras en documentos hablados. Los grafos de fonemas propuestos se construyen a partir de probabilidades fonéticas calculadas *frame* a *frame*. También se propone un modelo de error sobre los grafos de fonemas que permite simplificar los algoritmos de exploración que buscan secuencias fonéticas. Se aplica un modelo de duración de fonemas para reducir falsos positivos que penaliza adecuadamente la detección de secuencias fonéticas en intervalos temporales demasiado cortos.
Palabras clave: Grafos de Fonemas, Detección y Localización de Palabras, Programación Dinámica

Abstract: In this paper we propose the use of phoneme graphs for word spotting tasks. The proposed phoneme graphs are built using phonetic probabilities estimated at frame level. We also propose an error model on phoneme graphs which allows to simplify the exploration algorithms used for finding phonetic sequences. A phoneme duration model is also applied for avoiding the detection of too short phonetic sequences, which helps to reduce the number of false positive detections.
Keywords: Phoneme graphs, Word Spotting, Dynamic Time Warping

1. Introducción

La modelización y detección de unidades fonéticas es uno de los objetivos principales para muchas aplicaciones del ámbito de las tecnologías del habla. Aunque los sistemas de Reconocimiento Automático del Habla (RAH) se basan en las palabras como unidades lingüísticas básicas, y por tanto elementos del modelo de lenguaje, la modelización acústica de las palabras se representa en términos de las unidades básicas del habla, que son los fonemas (u otras unidades subléxicas similares). La limitación en el número de unidades de tipo fonético a partir de las cuales se pueden componer todas las palabras de un lenguaje, y la relación quasi-directa entre fonemas y letras, las convierten en las unidades ideales para representar las características acústicas de una lengua. Desgraciadamente, su corta duración y su variabilidad acústica hace muy difícil abordar tareas puras de Decodificación Acústico-Fonética (DAF). Además, las fuertes restricciones que se pueden imponer mediante un modelo de lenguaje basado en palabras, no son comparables con la débil estructura sintáctica que representaría las posibles concatenaciones de fonemas. Por ello el soñado objetivo que empezó a tomar fuerza a principios de los 80 (con el desarrollo de técnicas robustas de aprendizaje y modelización acústico-fonética, principalmente los basados en HMM), de disponer de sistemas que reconocieran directamente la secuencia de unidades fonéticas pronunciadas, y a partir de ellas obtener la secuencia de palabras, ha quedado algo olvidado por los métodos de Reconocimiento del Habla continua basado en palabras. No obstante, con el desarrollo de las tecnologías en los últimos años, y las nuevas necesidades creadas al abordar nuevas y más complejas tareas, la capacidad de detectar y reconocer unidades fonéticas en una pronunciación es una tarea que está tomando cada vez más interés.

* Los resultados presentados aquí son fruto del trabajo desarrollado dentro del proyecto TIN2008-06856-C05-02/TIN.

En la medida que se han abierto nuevas expectativas en el ámbito de las tecnologías del habla, como son el reconocimiento del habla para muy grandes vocabularios (Rastrow et al., 2009), los sistemas de recuperación y extracción de información en corpus hablados (Amir, Efrat, y Srinivasan, 2001; Saraclar y Sproat, 2004), o detección de palabras relevantes en este tipo de corpus, el reconocimiento a nivel fonético adquiere mayor importancia. Su aportación se basa en que en muchos casos es imposible tener representado todo el vocabulario y por tanto habrá que detectar palabras o subsecuencias léxicas en términos de secuencias fonéticas. Tal puede ser el caso de entidades nombradas, por ejemplo nombres propios en otros idiomas, que pueden ser usadas para detección y reconocimiento de palabras fuera del vocabulario o para extracción de información basada en subsecuencias léxicas (Ng y Zue, 1998).

En aproximaciones clásicas basadas en modelos de palabras, la búsqueda se circunscribe a encontrar palabras incluidas en el modelo de lenguaje utilizado para entrenar los sistemas de RAH, en cuyo caso resulta más eficiente localizar las palabras a buscar en la salida del reconocedor. Con esta aproximación se obtienen buenos resultados cuando se trabaja con las n-mejores frases que devuelve el reconocedor (Saraclar y Sproat, 2004). A partir de las aproximaciones basadas en modelos de palabras se construyen grafos de fonemas utilizando las secuencias fonéticas de las palabras de las n-mejores frases a la salida del reconocedor. Esto permite encontrar palabras fuera del vocabulario (Amir, Efrat, y Srinivasan, 2001).

En este trabajo se presenta una aproximación diferente, basada en grafos de fonemas para tareas de búsqueda de palabras en corpus de voz a partir de la secuencia fonética. Disponiendo de un transcriptor fonético se puede buscar cualquier palabra o secuencia de palabras que pida el usuario. Esto facilita la búsqueda de palabras que no pertenecen al vocabulario, y además permite la búsqueda de secuencias fonéticas cualesquiera. Secuencias fonéticas que pueden ser parte de una palabra o abarcar varias palabras, según sea de interés, por ejemplo, para la indexación de corpus de voz en base a secuencias fonéticas.

En el resto del documento se describe el sistema en la sección 2, se detallan los experimentos y se muestran los resultados en la sección 3, finalmente se presentan las conclusiones en la sección 4.

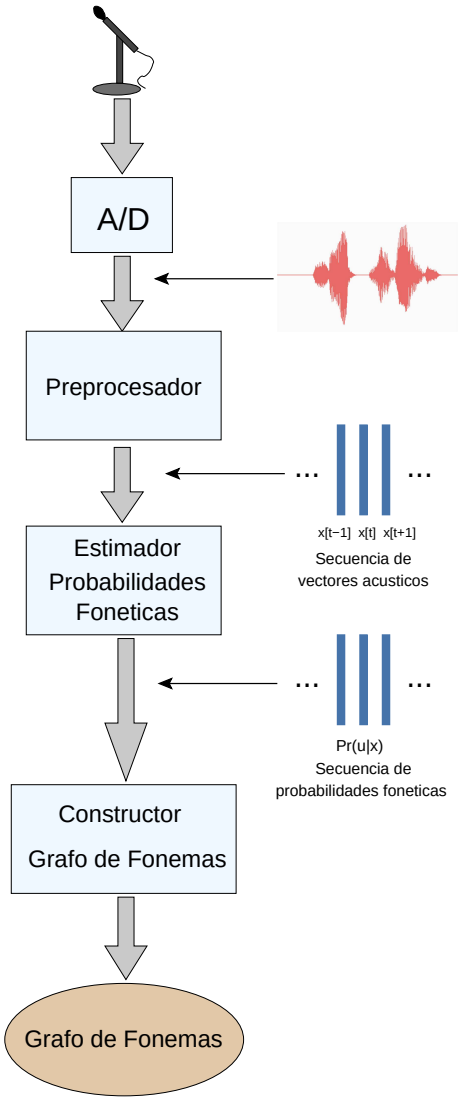

Figura 1: Sistema desacoplado para el procesado de la señal vocal por etapas

2. *Descripción del sistema*

Nuestro sistema para la detección y localización de palabras o secuencias fonéticas sobre documentos de audio es un sistema desacoplado, procesa la señal vocal por etapas. La Figura 1 muestra el esquema general con los módulos en disposición secuencial. Una implementación multihilo permite hacer eficiente un proceso desacoplado cuyos módulos funcionan en modo *pipe-line*. Este tipo de implementaciones adquiere especial relevancia desde la aparición de los microprocesadores multinúcleo. A continuación presentamos una breve descripción de las distintas etapas, cada una correspondiente a un módulo:

1. La señal vocal analógica capturada por el micrófono es digitalizada (típicamente a 16KHz) y filtrada.

2. El preprocesador convierte la señal vocal en una secuencia de vectores acústicos. En nuestro sistema cada vector acústico o *frame* contiene 39 parámetros: la Energía, los primeros 12 *Mel Frequency Cepstral Coefficents (MFCC)*, más las primeras y segundas derivadas. Se extrae una *frame* cada $10ms$ aplicando una ventana de Hamming de $20ms$.

3. La finalidad de los parámetros acústicos es distinguir las unidades fonéticas. Para ello el módulo extractor de probabilidades fonéticas convierte cada vector acústico en un vector de probabilidades fonéticas. La salida de este módulo es una secuencia de vectores con probabilidades fonéticas.

4. Explorando la evolución temporal de la probabilidad de cada unidad fonética localizamos segmentos en los que ha sido pronunciada. Determinando donde comienzan y donde acaban las distintas ocurrencias de cada unidad fonética se construye un grafo de fonemas como representación de una pronunciación.

2.1. Estimación de las probabilidades fonéticas

A diferencia de los sistemas de reconocimiento automático del habla (RAH) estándares, basados en modelos ocultos de Markov (HMM, *Hidden Markov Models*), nuestro sistema calcula las probabilidades "a posteriori" de que cada unidad fonética u haya sido pronunciada dado un vector acústico x_t, $Pr(u|x_t)$, donde el subíndice t representa el instante de tiempo. Esta aproximación permite obtener un vector con probabilidades fonéticas $\{Pr(u_i|x_t)\}\ i = 1..|U|$ a partir de cada vector acústico x_t, cumpliéndose $\sum_{u \in U} Pr(u|x_t) = 1$, donde U representa el conjunto de unidades fonéticas. Aplicando este proceso a cada *frame* se obtiene una secuencia de vectores de probabilidades como representación de una pronunciación.

El cálculo de las probabilidades fonéticas "a posteriori" se realiza combinando probabilidades acústicas con probabilidades condicionales estimadas al efecto.

2.1.1. *Clustering* a nivel acústico

Las probabilidades acústicas se obtienen a partir de una mixtura de Gaussianas o GMM (*Gaussian Mixture Model*), donde cada Gaussiana representa una clase acústica natural. El GMM se obtiene mediante un proceso de *clustering* paramétrico a nivel acústico, donde la estimación de los parámetros que definen las Gaussianas o distribuciones normales se obtienen por máxima verosimilitud en modo no supervisado (Duda, Hart, y Stork, 2001).

La idea subyacente a esta aproximación se basa en considerar que, una vez transformada la señal vocal en vectores acústicos d-dimensionales, éstos se distribuyen en una región del espacio $\mathbb{R}^d$, acumulándose en subregiones más densas según rasgos acústico-fonéticos similares. Las subregiones más densas obedecen a las distintas manifestaciones acústicas del tracto vocal. Entendemos que cada unidad fonética tiene muchas manifestaciones acústicas posibles, debidas, entre otros fenómenos, al estado de ánimo y al acento del locutor. También de manera destacada debidas al contexto, es decir, que la manera en que se pronuncia un fonema depende mucho de los fonemas que le preceden y de los que le suceden. Además, no todas las manifestaciones acústicas posibles pertenecen a un sólo fonema, hay muchas que se ubican en la intersección entre dos o más fonemas. Las subregiones de cada fonema no son estancas ni continuas.

Resumiendo, consideramos que las unidades fonéticas se distribuyen en subregiones solapadas dentro de $\mathbb{R}^d$, y que las clases acústicas naturales permiten modelar de manera más precisa la región de $\mathbb{R}^d$ por la que se distribuyen los vectores acústicos. Obviamente el número de clases acústicas será mucho mayor que el de unidades fonéticas.

2.1.2. Cálculo de las probabilidades fonéticas

Para tomar en consideración los distintos grados de relación entre las clases acústicas y las unidades fonéticas hemos utilizado probabilidades condicionales estimadas al efecto. Así, por cada unidad fonética disponemos de $Pr(a|u)$, es decir, la probabilidad de que la clase acústica a se haya manifestado cuando la unidad fonética u ha sido pronunciada. Estas probabilidades condicionales están normalizadas de manera que se cum-

ple $\sum_{a \in A} Pr(a|u) = 1$, donde A representa el conjunto de clases acústicas naturales.

En base al GMM podemos calcular $p(x_t|a)$, la densidad de probabilidad condicional de observar el vector acústico x_t cuando se ha manifestado la clase acústica a. Cuya fórmula es la conocida distribución normal o de Gauss. Combinando éstas con las probabilidades condicionales arriba mencionadas podemos calcular $p(x_t|u)$, la densidad de probabilidad condicional de observar el vector acústico x_t cuando ha sido pronunciada la unidad fonética u (Gómez y Castro, 2002):

$$p(x_t|u) = \sum_{a \in A} p(x_t|a) \cdot \Pr(a|u) \qquad (1)$$

y aplicando la regla de Bayes obtenemos las probabilidades fonéticas "a posteriori":

$$\Pr(u|x_t) = \frac{p(x_t|u)\pi(u)}{\sum_{v \in U} p(x_t|v)\pi(v)} \qquad (2)$$

donde $\pi(\cdot)$ es la probabilidad "a priori" de cada unidad fonética. En nuestro caso consideramos todas las unidades fonéticas equiprobables a priori, por tanto eliminando éstas y expandiendo la formulación de $p(x_t|u)$ obtenemos:

$$\Pr(u|x_t) = \frac{\sum_{i=1}^{|A|} p(x_t|a_i) \cdot \Pr(a_i|u)}{\sum_{j=1}^{|U|} \left(\sum_{i=1}^{|A|} p(x_t|a_i) \cdot \Pr(a_i|u_j) \right)} \qquad (3)$$

Las probabilidades condicionales $Pr(a|u)$ pueden estimarse bien de manera supervisada a partir de un corpus segmentado y etiquetado manualmente, bien de manera no supervisada mediante un proceso de refinamiento sucesivo de las fronteras fonéticas. Para el sistema presentado aquí estas probabilidades condicionales se estiman de manera no supervisada, dado que también es utilizado para segmentar y etiquetar automáticamente corpus de voz (Gómez y Castro, 2002).

2.2. Grafos de Fonemas

Los grafos de fonemas resultan adecuados para representar frases pronunciadas entre silencios, incluso cuando se trata de una sóla palabra. El hecho de considerar varias unidades fonéticas alternativas para un segmento de tiempo permite reflejar la incertidumbre asociada al proceso de detección fonética sin utilizar conocimiento de niveles superiores: léxico, sintáctico y semántico.

La Figura 2 muestra un ejemplo de los grafos de fonemas utilizados aquí, formados por nodos que representan instantes de tiempo y arcos etiquetados con unidades fonéticas. Los arcos también incorporan una medida de confianza, calculada como la probabilidad media de la unidad fonética desde la *frame* correspondiente al nodo origen hasta la *frame* anterior al nodo destino. Los nodos están etiquetados con el índice de vector acústico. Debemos tener presente que el módulo preprocesador emite un vector acústico de manera constante cada cierto intervalo de tiempo. Este intervalo de tiempo es invariable, típicamente $10ms$. Por tanto el índice representa el tiempo en intervalos de $10ms$.

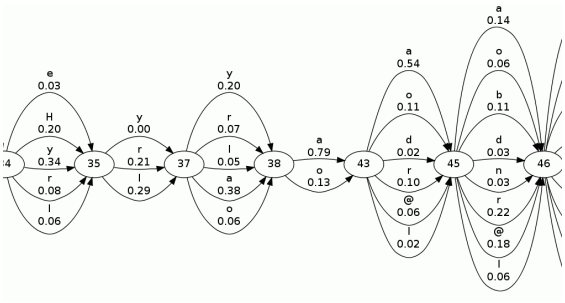

Figura 2: Trozo de un grafo de fonemas

La modalidad en que los nodos representan unidades fonéticas y los arcos representan las transiciones posibles entre éstas ha sido estudiada previamente por los autores (Gómez, Castro, y Sanchis, 2002).

2.2.1. Construcción

La construcción de un grafo de fonemas consiste en explorar la evolución temporal de las probabilidades fonéticas. Se trabaja con dos umbrales ajustados empíricamente, uno para detección y otro para ampliación. Por cada unidad fonética se observa si su probabilidad supera el umbral de detección, creándose entonces una hipótesis de segmento fonético que se extiende en ambos sentidos mientras se supera el umbral de ampliación. Una vez determinados los límites de un segmento fonético se crean dos nodos, uno para el instante inicial y otro para el instante final, en caso de que todavía no existiesen. Después se crea un arco entre cada dos nodos consecutivos desde el que representa el instante inicial hasta el que representa el instante final.

Los nodos se crean en aquellos instantes en

los que el proceso de detección de fonemas ha determinado que comienza o acaba cualquier unidad fonética. Los arcos siempre unen dos nodos consecutivos. Puede observarse en la Figura 2 cómo en segmentos temporales donde claramente se ha pronunciado un fonema hay pocos arcos: unidades `a` y `o` entre los nodos 38 y 43. Mientras que en otros segmentos más ambiguos hay más arcos, por ejemplo entre los nodos 45 y 46.

Las unidades fonéticas incluyen los fonemas del idioma a reconocer más el silencio `<SIL>` y la pausa corta `<SP>`.

2.2.2. Modelo de Error

Para cualquier tarea posterior como la construcción de grafos de palabras y la localización de palabras o segmentos de palabras, lo que se requiere es encontrar secuencias fonéticas dentro del grafo de fonemas. Al pretender completar secuencias fonéticas es muy habitual que éstas no existan completas, surgiendo la necesidad de permitir sustituciones, borrados e inserciones. El algoritmo que realiza estas operaciones necesita un modelo de error para estimar su coste.

Aquí se propone un modelo de error que consiste en añadir al grafo de fonemas los arcos que faltan, de manera que entre cada dos nodos consecutivos habrá un arco por cada unidad fonética. Estos nuevos arcos llevan asociada una medida de confianza, también relacionada con la probabilidad, que refleja el hecho de que la unidad fonética no se detectó originalmente. En la implementación de nuestro sistema para los experimentos aquí presentados hemos utilizado directamente las probabilidades fonéticas de las unidades a añadir, al igual como se han calculado las medidas de los arcos originales. Esto ya supone una adecuada penalización y aporta la ventaja de disponer de arcos para todas las unidades, en consecuencia el algoritmo que explora el grafo de fonemas es mucho más simple y eficiente, todas las operaciones son coincidencias.

Las medidas asignadas a los arcos añadidos por el modelo de error se pueden calcular mediante otras aproximaciones. Por ejemplo calculando una matriz de confusión con la probabilidad de que el sistema confunda una unidad fonética con otra, y utilizar estas probabilidades de confusión para calcular la medida de confianza de los arcos añadidos a partir de las medidas de confianza de los arcos originales. El estudio comparativo de diferentes variantes de modelo de error constituye una extensión al presente trabajo. El intercambio entre modelos de error es sencillo gracias a que la estructura de los grafos ya prevé utilizar distintos modelos de error.

En resumen, la incorporación de todos los arcos gracias al modelo de error facilita la búsqueda de secuencias fonéticas en el grafo de fonemas, donde todas las operaciones del algoritmo de búsqueda serán coincidencias. La penalización correspondiente ya está precalculada y reflejada en las medidas de confianza de los arcos añadidos.

2.2.3. Modelo de duración de fonemas

Al recorrer los grafos de fonemas para encontrar secuencias fonéticas, es habitual que éstas se localicen en intervalos excesivamente pequeños por el simple hecho de que el sistema ha detectado, de manera consecutiva, segmentos cortos de los fonemas que las componen, dando por buena una palabra en un intervalo donde es casi imposible que fuese pronunciada. Este fenómeno se agrava para las secuencias más cortas.

Con el fin de penalizar la asignación de fonemas a segmentos excesivamente cortos se aplica un modelo de duración de fonemas. Cada vez que se asocia un fonema de la secuencia a encontrar con varios arcos consecutivos la probabilidad fonética de los arcos se pondera según el modelo de duración.

El modelo de duración de fonemas utilizado en este trabajo es independiente del contexto y ha sido estimado con la misma base de datos utilizada en los experimentos. Cabe destacar que su aplicación ha sido determinante para la mejora de resultados. Como se aprecia en los experimentos de DAF, aplicar un modelo de duración incontextual mejora los resultados significativamente. La tasa de fonemas acertados pasa del 42,8 % al 63,2 %.

Analizando con más detalle vemos que la desproporción se debe al aumento de borrados cuando no se aplica el modelo de duración, es decir, símbolos de más que aparecen en la secuencia fonética correspondiente al mejor camino. Al calcular la distancia de edición con respecto a la secuencia correcta para obtener la tasa de aciertos estos símbolos incrementan el número de operaciones de borrado.

2.2.4. Uso de los grafos de fonemas

Los grafos de fonemas son aptos para distintos fines, por ejemplo: 1) Decodificación Acústico Fonética (DAF), 2) localización de las distintas ocurrencias de una o más palabras en documentos hablados (*Word Spotting*), o de subsecuencias fonéticas que no correspondan a una palabra completa, y 3) construcción grafos de palabras.

Tareas de detección y localización de secuencias fonéticas se pueden abordar desde diferentes perspectivas, siendo la más aplicada la que se basa en la salida de un sistema de RAH, obteniéndose buenos resultados si se trabaja con el grafo de palabras que representa las n-mejores frases reconocidas (Saraclar y Sproat, 2004).

Utilizar sistemas de RAH tiene la limitación de no poder localizar palabras de fuera del vocabulario. Otra perspectiva interesante y bastante aplicada es la de buscar secuencias fonéticas, lo que permite encontrar cualquier palabra disponiendo de su secuencia fonética (Amir, Efrat, y Srinivasan, 2001). En esta modalidad resultan de aplicación los grafos de fonemas descritos en el presente trabajo, cuyos resultados se presentan en la sección 3.

3. Experimentación

Los grafos de fonemas obtenidos para representar una pronunciación pueden servir para diferentes propósitos. En el trabajo presentado aquí se proponen como paso intermedio para detectar y localizar palabras en archivos de audio.

La recuperación de información a partir de audio (SDR: *Spoken Document Retrieval*), junto a la indexación de archivos de audio según un conjunto de palabras clave, son dos de las tareas más utilizadas en los últimos años para aprovechar la información contenida en las grabaciones de audio acumuladas desde hace décadas (Garofolo, Auzanne, y Voorhees, 2000). La detección y localización de palabras o *Word Spotting* es una de las aproximaciones más utilizadas para ambas tareas.

Con el objeto de evaluar la idoneidad de los grafos de fonemas construidos a nivel acústico y fonético, es decir, sin hacer uso de información a niveles superiores: léxico, sintáctico o semántico, se presentan dos tipos de experimentos: DAF y *Word Spotting*.

El corpus de voz utilizado en este trabajo ha sido el corpus fonético de la base de datos *Albayzin* (Moreno et al., 1993).

	sin ML	**con ML**
sin MD	42.8 %	53.7 %
con MD	63.2 %	65.6 %

Cuadro 1: *Tasa de aciertos a nivel de fonema al utilizar los grafos de fonemas para Decodificación Acústico-Fonética, con y sin modelo de lenguaje a nivel fonético combinado con la utilización o no de modelo de duración.*

3.1. DAF

Los experimentos de DAF tienen por objeto evaluar la capacidad de los grafos de fonemas anteriormente descritos para representar pronunciaciones. Su construcción se basa en la simple detección de segmentos en los que el sistema considere que una unidad fonética está presente. Obviamente no podemos esperar que los fonemas más probables coincidan con los realmente pronunciados. Es por ello que entre cada dos nodos existen varios arcos, etiquetados con las unidades fonéticas más probables y con una medida de confianza asociada. En la práctica el mejor camino a través del grafo de fonemas nunca contendrá la transcripción correcta al 100 %.

El Cuadro 1 presenta los resultados de DAF a partir de grafos de fonemas sin los arcos añadidos por el modelo de error, sólo con los arcos detectados originalmente. Nótese la significativa influencia de un modelo de duración de fonemas. Los porcentajes mostrados corresponden a la tasa de aciertos a nivel de secuencia fonética detectada, no a nivel de *frame*, donde los resultados son mejores pero no sirven de referencia. La tasa de aciertos se calcula con la siguiente expresión:

$$100 \cdot \frac{A}{A + S + B + I} \tag{4}$$

donde A es el contador de aciertos, S el de sustituciones, B el de borrados e I el de inserciones. Estos contadores se obtienen al calcular la distancia de edición entre la secuencia fonética correspondiente el mejor camino y la de referencia.

Adicionalmente hemos calculado la tasa de aciertos obtenida al buscar en los grafos de fonemas la secuencia fonética de referencia: 78,64 %. Este valor mide la capacidad del sistema para detectar los fonemas pronunciados aunque no sean siempre los más

probables. Cuando únicamente se consideran los más probables se obtiene el resultado del Cuadro 1 correspondiente a la combinación **sin** modelo de lenguaje a nivel fonético y **sin** modelo de duración de fonemas.

El modelo de lenguaje a nivel fonético únicamente se ha utilizado en los experimentos de DAF para ilustrar su influencia, pero no ha sido utilizado en los experimentos de la siguiente subsección.

3.2. *Word Spotting*

En nuestro sistema la detección y localización de palabras se lleva a cabo mediante un algoritmo de Programación Dinámica que intenta ubicar una secuencia fonética dentro de un grafo de fonemas. Este algoritmo utiliza los arcos originales más los añadidos por el modelo de error, luego no necesita contemplar posibles inserciones, borrados o sustituciones. También permite que la secuencia fonética a encontrar comience en cualquier nodo, no obliga a que la secuencia parta del nodo inicial y llegue hasta el nodo final. Gracias a esta característica el algoritmo puede encontrar varias ocurrencias de una misma secuencia fonética en una sola pasada.

El sistema construye un grafo de fonemas por cada pronunciación, entendiendo por pronunciación un segmento de voz entre dos silencios suficientemente largos. Por tanto, una pronunciación puede ser una palabra suelta, una frase o un párrafo. Sobre cada grafo de fonemas se buscan las palabras a localizar más todas las del vocabulario disponible. Incluir en la búsqueda todas las palabras de un amplio vocabulario permite podar para controlar la tasa de falsos positivos.

Para medir el rendimiento de nuestro sistema a nivel de *Word Spotting* hemos utilizado las medidas estándar *"Recall"* y *"Precision"*: Por cada palabra w se calculan como:

$$Recall(w) = \frac{Correctas(w)}{Referencia(w)} \qquad (5)$$

$$Precision(w) = \frac{Correctas(w)}{Detectadas(w)} \qquad (6)$$

donde $Correctas(w)$ representa el contador de aciertos, $Referencia(w)$ el número de ocurrencias realmente pronunciadas, y $Detectadas(w)$ el número de todas las veces que el sistema ha detectado w. Después se calculan los valores medios de *"Recall"* y *"Precision"* considerando todas las palabras utilizadas para test. En los experimentos aquí mostrados el conjunto de palabras a buscar es un subconjunto del vocabulario formado por las superiores a 6 fonemas.

La Figura 3 presenta una gráfica DET (*Detection Error Trade-off*), donde la diagonal representa el punto EER (*Equal Error Rate*). Este tipo de gráficas ilustra el comportamiento del sistema según se ajusta un umbral para encontrar un equilibrio entre tasa de aciertos y de falsos positivos. En nuestro caso el umbral es el límite inferior de la medida de confianza asociada a cada ocurrencia detectada para decidir si se acepta.

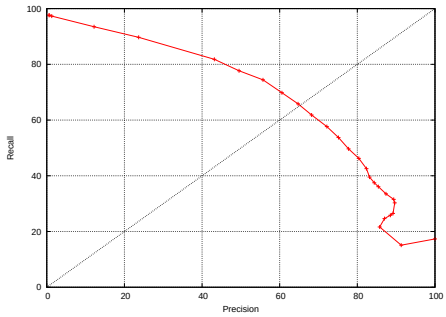

Figura 3: *Recall* frente a *Precision* sin poda.

En la Figura 3 se observa como el precio de obtener un alto porcentaje de aciertos es un excesivo número de falsos positivos (baja precisión). No obstante, el EER se sitúa cercano al 65 %, valor similar, y en algunos casos superior, a los EER obtenidos en otros trabajos mediante aproximaciones que utilizan conocimiento léxico y sintáctico pero con diferentes corpus de voz (Amir, Efrat, y Srinivasan, 2001), (Saraclar y Sproat, 2004).

Gracias a que la búsqueda se realiza para todas las palabras del vocabulario se puede podar para mejorar la precisión. La Figura 4 presenta el caso en que se han considerado las N-mejores palabras detectadas *frame* a *frame* (N=4). Esto quiere decir que una palabra detectada no ha sido desechada si en alguna de las *frames* que abarca ha estado entre las N-mejores. La Figura 5 ilustra el efecto de la poda para distintos valores de las N-mejores. Podar mejora el EER, consiguiendo situarlo en torno al 70 % para valores de N alrededor de 4.

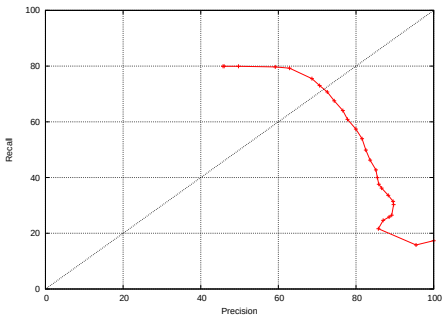

Figura 4: *Recall* frente a *Precision* considerando las 4 mejores.

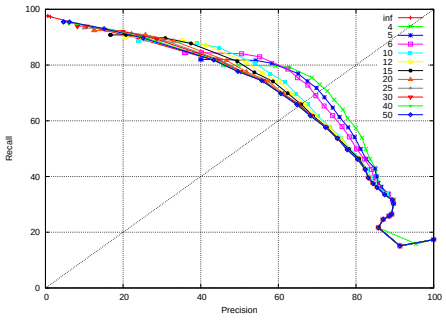

Figura 5: *Recall* frente a *Precision* para varios valores de poda. *inf* representa no poda.

4. Conclusiones

Hemos presentado la construcción de grafos de fonemas utilizando únicamente conocimiento acústico y fonético, y su aplicación a tareas de búsqueda y localización de palabras en documentos hablados.

Respecto de la construcción de los grafos de fonemas es destacable su sencillez y su bajo coste computacional. Del modelo de error es importante destacar el efecto que tiene para los algoritmos de exploración añadir los arcos que faltan, en el presente trabajo para el algoritmo utilizado en los experimentos de *Word Spotting*. Queda pendiente contrastar diferentes aproximaciones de modelo de error en trabajos futuros.

Otro aspecto a destacar es la aplicación de un modelo de duración de fonemas, para *Word Spotting* consigue una reducción considerable de falsos positivos gracias a que penaliza adecuadamente la detección de secuencias fonéticas en segmentos temporales demasiado cortos. Dada la mejora que también aporta a nivel de DAF el modelo de duración de fonemas incontextual, queda pendiente como desarrollo futuro el estudio de modelos de duración dependientes del contexto.

Bibliografía

Amir, Arnon, Alon Efrat, y Savitha Srinivasan. 2001. Advances in phonetic word spotting. En *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, páginas 580–582, New York, NY, USA. ACM.

Duda, R. O., P. E. Hart, y D. G. Stork. 2001. *Pattern Classification*. John Wiley and Sons, second edición.

Garofolo, John, Cedric G. P. Auzanne, y Ellen M. Voorhees. 2000. The trec spoken document retrieval track: A success story. En *Text Retrieval Conference (TREC) 8*, páginas 16–19.

Gómez, J.A. y M.J. Castro. 2002. Automatic Segmentation of Speech at the Phonetic Level. En *Structural, Syntactic, and Statistical Pattern Recognition*, volumen 2396 de *LNCS*. Springer-Verlag, August, páginas 672–680.

Gómez, J.A., M.J. Castro, y E. Sanchis. 2002. Construcción de grafos de fonemas para un sistema de RAH desacoplado. En *II Jornadas en Tecnología del Habla, Granada, España*, Granada, Spain, December.

Moreno, A., D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, y C. Nadeu. 1993. Albayzin Speech Database: Design of the Phonetic Corpus. En *Proc. Eurospeech93*, volumen 1, páginas 653–656, Berlin, Germany, September.

Ng, K. y V. Zue. 1998. Phonetic recognition for spoken document retrieval. En *ICASSP*, páginas 325–328.

Rastrow, Ariya, Abhinav Sethy, Bhuvana Ramabhadran, y Frederick Jelinek. 2009. Towards using hybrid word and fragment units for vocabulary independent lvcsr systems. En *Proceedings of the 10th Annual Conference of the International Speech Communication Association*.

Saraclar, Murat y Richard Sproat. 2004. Lattice-based search for spoken utterance retrieval. En *HLT-NAACL 2004: Main Proceedings*, páginas 129–136, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.